A texture-based probabilistic approach for lung nodule segmentation

Olga Zinoveva[1], Dmitriy Zinovev[2], Stephen A. Siena[3], Daniela S. Raicu[2], Jacob Furst[2], Samuel G. Armato[2]

[1] Harvard University, 200 Quincy Mail Ctr., Cambridge, MA
[2] College of Computing and Digital Media, DePaul University, 243 S. Wabash Ave, Chicago, IL 60604
[3] University of Notre Dame, Notre Dame, IN 46556
[2] Comprehensive Cancer Center, The University of Chicago 5841 South Maryland Avenue, MC 1140, Chicago, IL 60637

Abstract. Producing consistent segmentations of lung nodules in CT scans is a persistent problem of image processing algorithms. Many hard-segmentation approaches are proposed in the literature, but soft segmentation of lung nodules remains largely unexplored. In this paper, we propose a classification-based approach based on pixel-level texture features that produces soft (probabilistic) segmentations. We tested this classifier on the publicly available Lung Image Database Consortium (LIDC) dataset. We further refined the classification results with a post-processing algorithm based on the variability index. The algorithm performed well on nodules not adjacent to the chest wall, producing a soft overlap between radiologists' based segmentation and computer-based segmentation of 0.52. In the long term, these soft segmentations will be useful for representing the uncertainty in nodule boundaries that is manifest in radiological image segmentations.

Keywords: segmentation, probabilistic, lung, classifier, LIDC

1 Introduction

Most lung cancer treatment methods rely on early detection of malignant tumors. An effective way of measuring the malignancy of a lung nodule is by taking repeated computed tomography (CT) scans at intervals of several months and measuring the change in the nodule's volume[1]. However, the process of segmenting the nodule consistently is challenging, both for human readers and automated algorithms.

One of the greatest difficulties facing automatic lung nodule segmentation algorithms is the absence of a reliable and unambiguous ground truth. Many algorithms are trained on data from the Lung Image Database Consortium (LIDC)[2], which provides a reference truth based on the contours marked by four radiologists. Armato et al. explored the possible reference truths that may be constructed from the

sets of nodules detected by different radiologists on the same CT scans, and found significant variations[3]. This alone can greatly affect the results of a detection algorithm. The same is true of segmentation. Siena et al. measured the variability of radiologist contours in LIDC data and found that there are certain images for which disagreement is extremely high[4]. In such cases, it may be difficult to find a reliable reference truth.

The vast majority of lung nodule segmentation algorithms in the past have produced hard (binary) segmentations. Many methods for this kind of classification exist, but it is difficult to compare their effectiveness because they are quantified differently. For instance, Liu et al used the popular level set technique, though they did not provide an overall quantitative measure of their algorithm's accuracy[5]. Demeshki et al employed region growing and fuzzy connectivity and evaluated segmentation results subjectively with the help of radiologists[6]. Xu et al used dynamic programming to segment nodules with radiologist-defined seed points, though they did not test their algorithm on a dataset[7]. Q. Wang et al's and publication on dynamic programming[8] and J. Wang et al's paper on a 3D segmentation algorithm[9] evaluated their results by calculating the overlap of computer-generated segmentations against a ground truth. Q. Wang et al obtained overlaps of 0.58 and 0.66 on two datasets, and J. Wang et al had overlaps of 0.64 and 0.66, on two different datasets. Comparison of different methods is further complicated by the use of different datasets and varying methods of constructing the reference truths. The variation in the reference truth, which is usually produced by experienced radiologists, indicates that there may be more than one way to correctly segment lung nodules, so a probabilistic (soft) segmentation, which preserves variation in the data, may be a more natural way to segment nodules.

Soft segmentation has been applied to different areas of medical image processing, including segmentations of the kidneys[10] and magnetic resonance images of the brain[11]. However, little work has been done on soft segmentations of lung nodules. Ginneken produced a soft segmentation of the 23 nodules in the first version of the LIDC dataset using a region growing technique[12]. In this paper, we propose a new method for soft lung nodule segmentation that investigates the power of texture-based image features in segmenting lung nodules.

2 Materials and Methods

Our approach was to train a classifier using texture and intensity features, and then use it to classify pixels of interest. After this initial segmentation, we refined our results using a post-processing algorithm (Variability Index (VI)[4] Trimming) that trims those portions of the segmentation that appear to create the most variation in the data. In the next five sections, we explain our proposed soft segmentation approach.

2.1 LIDC Dataset, Probability Maps, and Data Preprocessing

The LIDC dataset is an expanding collection of CT scans analyzed at five US academic institutions in the effort to facilitate the testing of computer-aided diagnosis (CAD) systems. At the time of this study, the second version of the dataset, LIDC85, was available, containing 60 series of chest CT scans representing 149 nodules. Each scan was presented separately to 4 radiologists, who provided contours for all nodules they found between 3 and 30 mm. Each nodule, therefore, was outlined by up to 4 radiologists. Given that we are investigating a soft segmentation approach, and therefore several boundaries per slice were needed to train and test our approach, we created our dataset as a subset of the LIDC85: 39 nodules on 326 images, selected based on the criterion that each would contain at least one 2D slice with 4 contours; 264 were in this category. The other 62 images were the remaining slices for the 39 nodules that contained fewer than 4 contours, as radiologist opinion may have differed with regard to the superior and inferior slices of a nodule.

The contours produced by the radiologists were translated into probability maps for analysis (Fig 1 A). In a probability map (p-map), each pixel of the image is assigned a probability of belonging to the structure of interest (a lung nodule, in our case). Since up to four radiologists annotated each nodule, each pixel can take on 5 discrete probability values (0, 0.25, 0.50, 0.75, and 1), depending on the number of radiologists that included that pixel within their contours. In our algorithms, these probability values were replaced with (0, 1, 2, 3, and 4), which indicate the number of radiologists that included the pixel.

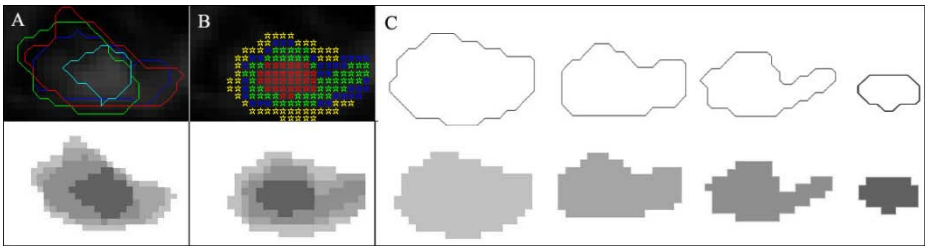

Fig. 1. Important terms (A) Radiologist outlines for a nodule in the LIDC dataset (top) and the corresponding p-map (bottom), where the level of probability is indicated by the color, from white (0) to the dark grey(1). (B) A computer generated p-map produced for the nodule in A using a decision tree classifier, overlaid on the CT scan (top) and displayed similarly to the radiologist p-map (bottom). (C) Thresholded p-maps for the computer-generated p-map in B (bottom), 0.25 (leftmost) to 1 (rightmost), and their corresponding contours (top).

Due to the varying properties and settings of the different scanners used to collect the LIDC data, pixel intensity histograms were not consistent on a series-to-series basis. At least four brands of scanners were used, including ones from GE, Toshiba, Siemens, and Philips, and certain settings and data display options were highly variable. The most significant variable that defined the histograms was the rescale intercept b used in the series of scans. In order to make intensity values comparable

across images, we modified all intensity histograms by shifting their rescale intercept to -1024 (the most common value).

When training the classifier, we selected 10,000 pixels for the nodule class and 10,000 pixels for the non-nodule class. The pixels were selected from the radiological p-maps. Every pixel was assigned a value 0-4, indicating the number of outlines that included it, so that pixels selected by four radiologists would have a value of four. Training pixels were selected only from those slices of the 39 nodules that contained outlines from four radiologists.

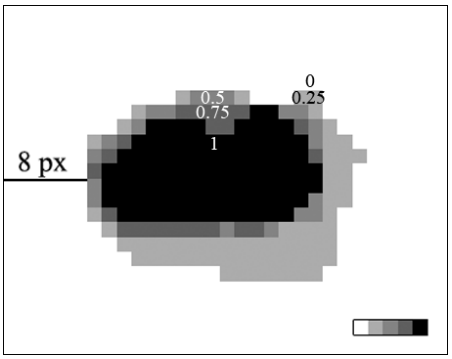

Fig. 2. An example of a p-map constructed from radiologists' outlines of a nodule. The shades of gray represent probabilities 0, 0.25, 0.5, 0.75, and 1

In the selection of random points, the non-nodule pixels were those that lay outside of the p-map, but inside a rectangular box that included the nodule and added 8 pixels in all four directions (Fig. 2). This 8-pixel box was selected because the high running time of certain feature calculation algorithms did not allow for a larger one, and anything smaller would have prevented us from evaluating the algorithm's performance on structures surrounding the nodule. The nodule pixels were selected from the 0.25, 0.50, 0.75, and 1 p-map areas.

2.2 Feature Extraction

Once random training pixels were selected, we performed feature extraction on the pixel level. We calculated the following features in a 9x9 neighborhood around the pixel of interest: intensity (including the intensity of the pixel of interest, as well as the minimum, maximum, mean, and standard deviation of the intensities in the neighborhood), Gabor filters, and Markov Random Fields.

To extract Gabor and Markov features, we used an open-source implementation of a feature extractor called BRISC[13]. Gabor filters are harmonic functions modulated by a Gaussian function. As per the algorithm, 12 Gabor filters were used: combinations of four orientations (0, $\pi/4$, $\pi/2$, $3\pi/4$ θ) and three frequencies (0.3, 0.4, 0.5 $1/\lambda$). A Markov random field is a matrix of random variables that exhibit a Markov property with respect to their neighbors. In the BRISC implementation, four Gaussian Markov random field parameters (corresponding to four orientations

between two neighboring pixels – 0^o, 45^o, 90^o, 135^o) and their variance were calculated.

2.3 Decision Tree Classification

We built our classifier using the Classification and Regression Trees (C&RT) binary decision tree provided by SPSS. We trained the classifier on 10,000 non-nodule and 10,000 nodule pixels, which represented 3 percent of the entire pixel dataset. We then used the result to classify all the 525,084 pixels that lay within the 8-pixel offset box described in Fig 2. The classifier assigned each pixel a continuous probability (*CP'*) from 0 to 1 of belonging to the nodule. We constructed computer-generated p-maps (Fig. 1 B) by binning these probabilities to make them comparable to radiologist p-maps during analysis of the data. We found discrete probability values *CP* as follows:

$$CP = \begin{cases} 1 \; if \; CP' > \dfrac{\overline{CP}_1 + \overline{CP}_{0.75}}{2} \\ 0.75 \; if \; CP' > \dfrac{\overline{CP}_{0.75} + \overline{CP}_{0.5}}{2} \\ 0.5 \; if \; CP' > \dfrac{\overline{CP}_{0.5} + \overline{CP}_{0.25}}{2} \\ 0.25 \; if \; CP' > \dfrac{\overline{CP}_{0.25} + \overline{CP}_0}{2} \\ 0 \; if \; CP' \leq \dfrac{\overline{CP}_{0.25} + \overline{CP}_0}{2} \end{cases}, \qquad (1)$$

where $\overline{CP}_{(1,0.75,0.50,0.25,0)}$ is the average probability assigned by the decision tree to all training set pixels originating from the respective p-map area and CP' is the probability value assigned by the classifier to the specific pixel in the test set. For instance, if the average probability assigned by the decision tree to the pixels originating from p-map area 1 is 0.92, and the average probability assigned to those originating from p-map area 0.75 is 0.76, any pixel above 0.84 would be assigned the value of 1 on the p-map by the algorithm. We attempted different methods for finding the thresholds, including hard-coding values, but we have found that this approach performs best.

2.4 Post-processing

To improve the results of our initial segmentation, especially with regard to systematic mistakes and over-segmentation, we used a post-processing algorithm called VI Trimming, which requires a seed point selected from the p-map generated by the classifier.

First, we constructed thresholded p-maps (Fig. 1C) of our soft segmentations for each of the probability areas (when a scan contained multiple nodules, each of the segmentations was treated separately when finding seed points). These p-maps were

then passed through a built-in Matlab implementation of a Savitzky-Golay Filter[14] This filter reduces the impact of noise in an image by moving a frame of a specified size over each column of an image and performing a polynomial regression on the pixels in that frame. The value of each pixel in the frame is then replaced by its value as predicted by the polynomial. We used a filter with a polynomial order of 3 and frame size of 7.

After noise reduction, we produced contours of each of the thresholded p-maps and found the centroid of the most circular contour on the highest-valued p-map, ignoring noise. This centroid was the seed point for the image. The VI Trimming algorithm began with the seed point for the image, then iteratively increased the area around it, starting with a 3x3 square, then growing to 5x5, etc. This square was filled with the computer-generated p-map. For each p-map square, a variability matrix was calculated according to the algorithm developed by Siena *et al*[4].

Once calculated, the variability matrix was used to generate a pointer matrix. The pointer matrix is a border that surrounds the outer edge of the variability matrix. Each pixel in the pointer matrix indicates how many variability matrix pixels adjacent to it are below the selected VI threshold. We have found 2 to be the optimal VI threshold value for our purpose. All pixels outside of a 0 in the pointer matrix are reset to 0 in the post-VI Trimming p-map, regardless of their original probabilities. This ensures that the nodule region that contained the seed points is kept in the p-map, while other regions are eliminated. This is the key step for removing misclassified chest wall regions, given that they are separated by at least a small gap from the nodule itself. The matrix stops growing when all the pixels in the pointer matrix are 0.

2.5 Evaluation of the Segmentation

We evaluated the quality of our segmentations using two metrics: the soft overlap[12] and the variability index[4] (not to be confused with VI Trimming, which is based on the variability index). The soft overlap is a measure of agreement between two soft segmentations. Values range between 0 for completely dissimilar segmentations, and 1 for identical segmentations. In our case, we compare our computer-generated p-maps against radiologist-generated ones.

It is calculated as follows:

$$SO = \frac{\sum_{(i,j)} \min(RP_n(i,j), CP_n(i,j))}{\sum_{(i,j)} \max(RP_n(i,j), CP_n(i,j))} \forall n \epsilon I \qquad (2)$$

where $CP_n(i,j)$ is the p-map value for pixel (i,j) in the computer-generated p-map of image n, and $RP_n(i,j)$ is the p-map value for the same pixel in the radiologist-generated p-map.

The variability index is a metric for evaluating the variability of a soft segmentation. Given a probability map, it takes into account both the number of pixels with probability below 1, and the shape of each probability area. We used the method described in the VI Trimming section to find the variability matrix for each image. Then, we calculated the variability index *VI* for each image[4]:

$$VI = \frac{\sum_{(i,j)} V(i,j)}{\sum_{(i,j)} P_{(i,j)}} \tag{3}$$

where P(i,j) denotes RP(i,j) or CR(i,j), depending on whether the VI is calculated for radiologist or computer segmentations.

3 Results

We used cross-validation on the training set and obtained the lowest risk estimate (variance about the mean of the node) of 0.06 for a decision tree with a depth of 10 and 53 terminal nodes. Before post-processing, our classifier had a median SO of 0.49 on the 39 nodules in our subset of the LIDC85 (Fig. 3A). The variability index distribution was highly right-skewed, indicating a few outliers with very high variability (Fig. 3B). The median VI was 4.50. Using the Inter-quartile range criterion, there were 42 possible outliers, specifically all images with VI above 16.88.

To improve these results, we used VI Trimming. The median SO rose to 0.52 (Fig. 3C). The large number of nodules with an SO lower than 0.1 are a result of misclassification on specific groups of nodule slices. Specifically, the classifier did not perform well on superior and inferior slices of each nodule, and nodule slices in contact with the chest wall. Furthermore, a failure to select good seed points resulted in lower post-VI Trimming SO for certain nodule slices.

After VI Trimming, the median variability index for all images decreased to 2.61. There were 20 outliers, which were all images with VI above 12.09 (Fig. 3D). For examples of VI Trimming results, refer to Fig. 4.

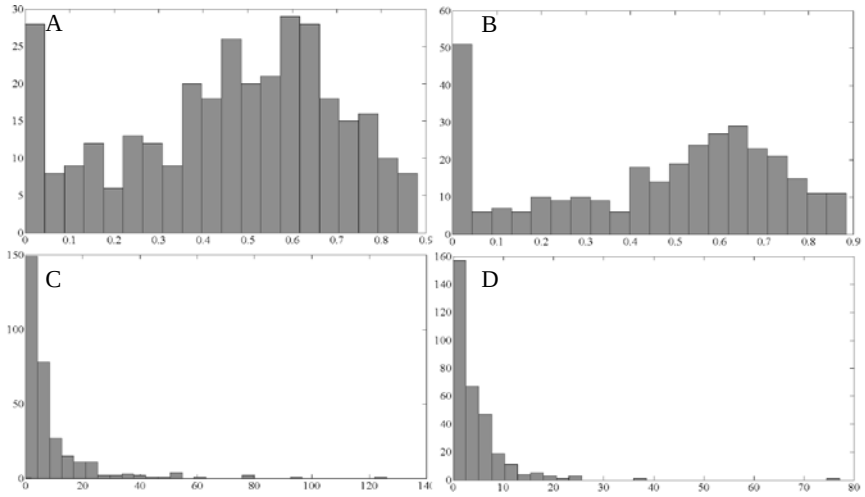

Fig. 3. Soft overlaps and variability indices before (A, B) and after (C, D) VI Trimming

In addition to calculating the variability index for all images, we found it specifically for those that were post-processed with VI Trimming (221 images). For

these images, the median VI decreased from 3.63 to 2.01. More significantly, the number of outliers in this case decreased from 28 to 18, indicating that VI Trimming is useful for minimizing the number of highly variable outliers.

For a summary of the results, see Table 1.

Table 1. Summary of results before and after post-processing

	Algorithm only	Algorithm plus VI Trimming
Soft Overlap	0.49	0.52
A ratio for 0.25-thresholded p-map	0.48	0.55
Variability Index	4.50	2.61

4 Discussion

Our results indicate that decision tree classifiers trained on Gabor, Markov, and intensity image features can be used to produce soft segmentations of lung nodules. The classifier successfully distinguished nodules from adjacent blood vessels (Fig. 4 E-F) in the majority of cases, but it failed to differentiate between chest wall and nodule pixels (Fig. 4 A, C). The best way to improve upon these results is to include lung segmentation in the pre-processing step, which we plan to do in future work. Once lung segmentation is performed, we will also run our algorithm on all pixels within the lung, instead of only ones inside the 8-pixel offset box.

We compared our soft segmentation results against another soft lung nodule segmentation algorithm. Ginneken's region-growing algorithm produced a mean soft overlap of 0.68 for 23 nodules. Although the soft overlaps for these nodules are higher than ours, the algorithm described in Ginneken's paper included a pre-processing lung segmentation step, which makes the comparison more difficult. Additionally, Kubota et al's work on nodule segmentation shows a decrease in performance in moving from the first to the second LIDC dataset, on which they obtained 0.69 and 0.59 mean overlaps, respectively[15]. Kubota et al conclude that this is due to the second dataset's thicker slices and some subtle nodules, which also complicates the comparison of our results with those of Ginneken.

Due to the possible bias of selecting training pixels from the same nodules that the testing set came from, we also applied our algorithm to four additional nodules from the most recent LIDC dataset. These nodules were not involved in producing the classifier, so they ensured that there was no such bias when they were segmented. We obtained an SO of 0.44 for these nodules. This may have resulted from the higher variability in scanning methods in the most recent LIDC, which results in more variable data. In the future, we plan to run our algorithm on this version of the LIDC and investigate the change in results.

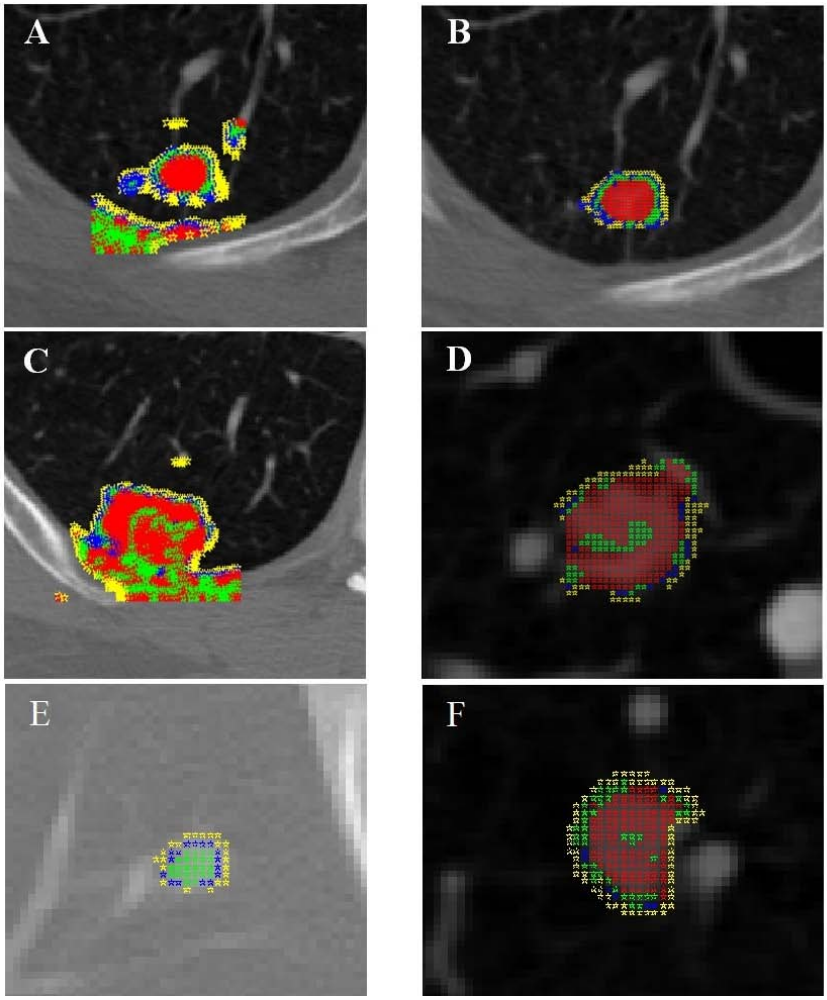

Fig. 3. Soft segmentations before and after VI Trimming. (Yellow – 0.25, blue – 0.5, green – 0.75, red – 1). (A) A nodule next to the chest wall before post-processing. The classifier could not distinguish the chest wall from the nodule. (B) The same nodule after VI Trimming. The chest wall has been removed. (C) VI Trimming was unable to correct the classifier's errors in this case because the nodule is too close to the chest wall. (D) The VI Trimming had no effect on this nodule because the classifier produced a good segmentation. (E-F) The classifier is good at distinguishing nodules from blood vessels, even without VI Trimming.

References

1. Ko, J.P., Betke, M.: Chest CT: automated nodule detection and assessment of change over time--preliminary experience. Radiology 218, 267-273 (2001)

2. Armato, S.G., 3rd, McLennan, G., McNitt-Gray, M.F., Meyer, C.R., Yankelevitz, D., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., Reeves, A.P., Croft, B.Y., Clarke, L.P.: Lung image database consortium: developing a resource for the medical imaging research community. Radiology 232, 739-748 (2004).
3. Armato, S.G., 3rd, Roberts, R.Y., Kocherginsky, M., Aberle, D.R., Kazerooni, E.A., Macmahon, H., van Beek, E.J., Yankelevitz, D., McLennan, G., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Caligiuri, P., Quint, L.E., Sundaram, B., Croft, B.Y., Clarke, L.P.: Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of "truth". Acad Radiol 16, 28-38 (2009)
4. Siena, S., Zinoveva, O., Raicu, D., Furst, J., Armato III, S.: A shape-dependent variability metric for evaluating panel segmentations with a case study on LIDC, In: Karssemeijer, N., Summers, R.M. (Eds.), 1 ed. SPIE, San Diego, California, USA, pp. 762416-762418 (2010)
5. Liu, S., Li, J.: Automatic medical image segmentation using gradient and intensity combined level set method. Conf Proc IEEE Eng Med Biol Soc 1, 3118-3121 (2006)
6. Dehmeshki, J., Amin, H., Valdivieso, M., Ye, X.: Segmentation of pulmonary nodules in thoracic CT scans: a region growing approach. IEEE Trans Med Imaging 27, 467-480 (2008)
7. Xu, N., Ahuja, N., Bansal, R.: Automated lung nodule segmentation using dynamic programming and EM-based classification, In: Sonka, M., Fitzpatrick, J.M. (Eds.), 1 ed. SPIE, San Diego, CA, USA, pp. 666-676 (2002)
8. Wang, Q., Song, E., Jin, R., Han, P., Wang, X., Zhou, Y., Zeng, J.: Segmentation of lung nodules in computed tomography images using dynamic programming and multidirection fusion techniques. Acad Radiol 16, 678-688 (2009)
9. Wang, J., Engelmann, R., Li, Q.: Computer-aided diagnosis: a 3D segmentation method for lung nodules in CT images by use of a spiral-scanning technique, In: Giger, M.L., Karssemeijer, N. (Eds.), 1 ed. SPIE, San Diego, CA, USA, pp. 69151H-69158 (2008)
10. Tang, H., Dillenseger, J.L., Bao, X.D., Luo, L.M.: A vectorial image soft segmentation method based on neighborhood weighted Gaussian mixture model. Comput Med Imaging Graph 33, 644-650 (2009)
11. Hongmin, C., Verma, R., Yangming, O., Seung-koo, L., Melhem, E.R., Davatzikos, C.: Probabilistic segmentation of brain tumors based on multi-modality magnetic resonance images, Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on, pp. 600-603 (2007)
12. van Ginneken, B.: Supervised probabilistic segmentation of pulmonary nodules in CT scans. Med Image Comput Comput Assist Interv 9, 912-919 (2006)
13. Lam, M.O., Disney, T., Raicu, D.S., Furst, J., Channin, D.S.: BRISC-an open source pulmonary nodule image retrieval framework. J Digit Imaging 20 Suppl 1, 63-71 (2007)
14. Savitzky, A., Golay, M.J.E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry 36, 1627-1639 (1964)
15. Kubota, T., Jerebko, A.K., Dewan, M., Salganicoff, M., Krishnan, A.: Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. Med Image Analysis 15, 133-154 (2011)